Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 193-196 recibido 13-12-2020 revisado 24-01-2021 aceptado 03-02-2021

Negation Processing in Spanish and its Application to Sentiment Analysis

Tratamiento de la Negación en Español y su Aplicación al Análisis de Sentimientos

Salud María Jiménez-Zafra
SINAI, Department of Computer Science, CEATIC, Universidad de Jaén,
Campus Las Lagunillas s/n, 23071, Jaén (Spain)
sjzafra@ujaen.es

Abstract: This is a summary of the Ph.D. thesis written by Salud María Jiménez Zafra at Universidad de Jaén under the supervision of Ph.D. María Teresa Martín Valdivia and Ph.D. L. Alfonso Ureña López. The author was examined on Friday, September 13th, 2019 by a committee formed by Ph.D. Ruslan Mitkov from the University of Wolverhampton, Ph.D. Miguel Ángel García Cumbreras from Universidad de Jaén and Ph.D. Eugenio Martínez Cámara from Universidad de Granada. The Ph.D. thesis obtained Summa Cum Laude and the international mention. Moreover, it was awarded as the best thesis in Natural Language Processing at the 36th International Conference of the Spanish Society for Natural Language Processing (SEPLN 2020).
Keywords: Negation processing, sentiment analysis, machine learning, natural language processing.

Resumen: Este es un resumen de la tesis doctoral realizada por Salud María Jiménez Zafra en la Universidad de Jaén bajo la dirección de los doctores Dña. María Teresa Martín Valdivia y D. L. Alfonso Ureña López. El acto de defensa tuvo lugar el viernes 13 de septiembre de 2019 ante el tribunal formado por los doctores D. Ruslan Mitkov de la Universidad de Wolverhampton, D. Miguel Ángel García Cumbreras de la Universidad de Jaén y D. Eugenio Martínez Cámara de la Universidad de Granada. La tesis obtuvo la calificación de Sobresaliente Cum Laude por unanimidad y mención de doctorado internacional. Además, recibió el premio a la mejor tesis doctoral en Procesamiento del Lenguaje Natural en el XXXVI Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2020).
Palabras clave: Tratamiento de la negación, análisis de sentimientos, aprendizaje automático, procesamiento del lenguaje natural.

1 Introduction

Natural Language Processing (NLP) is the field of Artificial Intelligence that aims to provide mechanisms to facilitate the communication between humans and machines through natural language (Indurkhya y Damerau, 2010). This is a challenge because computers must be able to process, understand and generate language. If we want to develop systems that approach human understanding, we must incorporate in them the processing of a diversity of linguistic phenomena, such as negation, irony or sarcasm, which are used to give words a different meaning.

This doctoral thesis focuses on the study of one of the main linguistic phenomena that we use in our daily communication, *negation* (Horn, 1989; Morante y Sporleder, 2012). Four tasks are usually performed in relation to processing negation: (i) negation cue detection, in order to find the words that express negation; (ii) scope identification, in order to find which parts of the sentence are affected by the negation cues; (iii) negated event recognition, to determine which events are affected by the negation cues; and (iv) focus detection, in order to find the part of the scope that is most prominently negated. Example (1) shows a sentence in which the negation cue appears in bold, the event in italics, the focus underlined and the scope between brackets.

ISSN 1989-7553. DOI 10.26342/2021-66-16

1. Es una persona que [**no** *tiene* límites], aunque a veces puede controlarse.
   He is a person who has no limits, although sometimes he can control himself.

This thesis addresses negation cue detection and scope identification tasks. In contrast to most of the studies existing so far that are on English language, it is carried out on Spanish texts, since not even Google has been able to process Spanish negation adequately. For example, if we perform in Google the search -*películas que no sean de aventuras*-, we can see that it returns adventure films when it should return films of other themes. Negation processing is not only important for information retrieval systems. In other systems, such as those of sentiment analysis, not processing negation can lead to the extraction of a completely different opinion than the one expressed by the user. For example, the polarity of the sentence "Una película fascinante, repetiría" *[A fascinating film, I would repeat]* should be the opposite of its negation "Una película nada fascinante, no repetiría" *[A not at all fascinating film, I would not repeat]*.

The objective of this dissertation is to advance in the processing of negation in Spanish and to show the importance of the computational treatment of negation for NLP systems. For this, an exhaustive study of negation is carried out, incorporating negation processing systems, corpora and sentiment analysis systems in which negation has been taken into account. In addition, a typology of negation patterns in Spanish is defined, which is applied for the annotation of a corpus with negation, the SFU Review$_{SP}$-NEG corpus. This corpus is used to develop a Spanish negation processing system which is applied to sentiment analysis in order to improve the predictive capacity of opinion classification systems that are so in demand today. Finally, NEGES has been launched, the first initiative promoting negation research in Spanish for which three editions have already been held in the context of the International Conference of the Spanish Society for Natural Language Processing (SEPLN).

2 Structure

This thesis is organized in eight chapters and one appendix, which are described hereafter.

Chapter 1 presents the motivation, objective and difficulty of the research addressed.

Chapter 2 introduces the concepts of negation and sentiment analysis, and presents the state-of-the-art for negation processing systems, the corpora annotated with negation, and sentiment analysis systems that take into account negation.

Chapter 3 shows the preliminary research, which reveals the importance of a correct processing of negation and the need to annotate a corpus with sentiment and negation. Spanish sentiment analysis systems existing up to now take negation into account as one more feature, but its effect on the classification is not evaluated.

Chapter 4 presents the SFU Review$_{SP}$-NEG corpus and the process followed for its annotation. In this chapter the components of negation are defined and delimited and it is proposed a typology of negation patterns in Spanish, which is applied for the annotation of the corpus. Moreover, it includes the annotation scheme used, the annotation process followed, the main sources of disagreement and the statistics and description of the corpus.

Chapter 5 includes all the details of the Spanish negation processing system developed. It contains an exhaustive analysis of the existing corpora in order to select the set of data for training the system. In addition, it presents the architecture of the proposed system, the experiments carried out, the results obtained and an analysis of errors aimed at understanding the limitations of the system.

Chapter 6 corresponds to the integration of the Spanish negation processing system developed into a sentiment analysis system. It presents the methodology followed to study the effect of negation, the experiments carried out and the results obtained, as well as an error analysis. It shows the importance of the development of accurate negation processing systems for NLP tasks.

Chapter 7 presents NEGES: Workshop on Negation in Spanish, the first initiative promoting negation research in Spanish. It contains the details of the origin of the workshop, its objective, the editions held, the tasks proposed, the datasets provided and the participants and results obtained.

Chapter 8 summarizes the conclusions, the main contributions, the research awards

and distinctions obtained, and the future lines of work.

Finally, **Appendix A** contains the tables summarizing the corpora analysis carried out in Chapter 5.

3 Contributions

Its main contributions can be grouped into 5 categories: state-of-the-art, resources, systems, analysis, and workshops.

State-of-the-art. We provided a thorough review of the work developed so far on the following topics (Jiménez-Zafra et al., 2018a; Jiménez-Zafra et al., 2019c): (i) English and Spanish negation processing systems; (ii) English and Spanish sentiment analysis systems that take into account negation; and (iii) Corpora annotated with negation.

Resources. Until now, there was no typology in Spanish to characterize and classify negation. Therefore, we defined our own (i) typology of negation patterns (Martí et al., 2016) taking into account their syntactic structure and their semantic interpretation. In addition, we defined an (ii) annotation scheme for negation and how it affects the sentiment of the sentence (Jiménez-Zafra et al., 2018b). We also generated (iii) the SFU Review$_{SP}$-NEG corpus (Jiménez-Zafra et al., 2018b) [1], the first corpus annotated with negation in the review domain for Spanish in which it is annotated how negation affects the words that are within its scope, that is, whether there is a change in the polarity or an increase or decrease of its value. Moreover, we presented (iv) a compilation of the corpora annotated with negation for all languages (Jiménez-Zafra et al., 2019c).

Systems. We developed (i) a polarity classification system for Spanish tweets that incorporates a set of syntactic rules for determining the scope of negation (Jimenez-Zafra et al., 2017). This rule-based approach has been proved to be better than the method most used to determine the scope of negation in English tweets. Furthermore, we implemented (ii) a machine learning system to process negation in Spanish (Jiménez-Zafra et al., 2020a). This system outperforms state-of-the-art results for negation cue detection, whereas for scope identification it is the first system that performs the task for Spanish.

[1]First Online: 22 May 2017 https://doi.org/10.1007/s10579-017-9391-x

Analysis. We reported (i) the problematic cases found during the annotation of the SFU Review$_{SP}$-NEG corpus in order to facilitate the annotation of this phenomenon for other researchers (Jiménez-Zafra et al., 2016). We also conducted (ii) an analysis of the corpora annotated with negation discussing the possibility of merging the corpora to create a larger data set to train a negation processing system. Moreover, we showed overall negation processing tasks for which the corpora could be used, and specific tasks for which the corpora could be used to evaluate the impact of processing negation. In addition, we provided (iii) a qualitative error analysis showing which negation cues and scopes are straightforward to predict automatically, and which ones are challenging (Jiménez-Zafra et al., 2020a). Furthermore, we studied (iv) the effect of the negation processing system developed on the sentiment analysis task (Jiménez-Zafra et al., 2020b).

Workshops. Finally, we created NEGES group and NEGES workshop, the first initiative promoting negation research in Spanish (Jiménez-Zafra et al., 2018a; Jiménez-Zafra et al., 2018b; Jiménez-Zafra et al., 2019a; Jiménez-Zafra et al., 2019b). NEGES is the acronym for "NEGación en ESpañol" (Negation in Spanish). It provides a means of exchanging news of recent research developments and other matters of interest as well as it makes available resources relevant to negation detection in Spanish, including corpora, annotation guidelines, evaluation scripts, etc. Up to now, three editions of NEGES have been held in the context of the SEPLN International Conference.

4 Conclusions

Negation is a complex linguistic phenomenon and the issue of its computational treatment has not been resolved yet due to its complexity, the multiple linguistic forms in which it can appear and the different ways it can act on the words within its scope. All languages possess different types of resources (morphological, lexical, syntactic) that allow speakers to speak about properties that people or things do not hold or events that do not happen. The presence of a negation in a sentence can have enormous consequences in many real world situations, for example, when processing clinical records. One might think that, given the fact that negations are

so crucial in language, most NLP pipelines incorporate negation modules and that the computational linguistics community has already addressed this phenomenon. However, this is not the case. Work on processing negation has started relatively late as compared to work on processing other linguistic phenomena. This doctoral thesis aims to advance the study of negation processing in Spanish.

Acknowledgements

This Ph.D. thesis has been partially supported by a grant from the Ministerio de Educación Cultura y Deporte (MECD - scholarship FPU014/00983), Fondo Europeo de Desarrollo Regional, LIVING-LANG project (RTI2018-094653-B-C21), REDES project (TIN2015-65136-C2-1-R) and ATTOS project (TIN2012-38536-C03-0) from the Spanish Government.

References

Horn, L. R. 1989. *A natural history of negation*. CSLI Publications.

Indurkhya, N. y F. J. Damerau. 2010. *Handbook of Natural Language Processing*, volumen 2. CRC Press.

Jiménez-Zafra, S. M., N. P. Cruz Díaz, R. Morante, y M. T. Martín-Valdivia. 2018a. Tarea 1 del Taller NEGES 2018: Guías de Anotación. En *Proceedings of NEGES 2018: Workshop on Negation in Spanish*, volumen 2174, páginas 15–21, Seville, Spain. CEUR-WS.

Jiménez-Zafra, S. M., N. P. Cruz Díaz, R. Morante, y M. T. Martín-Valdivia. 2018b. Tarea 2 del Taller NEGES 2018: Detección de Claves de Negación. En *Proceedings of NEGES 2018: Workshop on Negation in Spanish*, volumen 2174, páginas 35–41, Seville, Spain. CEUR-WS.

Jiménez-Zafra, S. M., N. P. Cruz Díaz, R. Morante, y M. T. Martín-Valdivia. 2019a. NEGES 2018: Workshop on Negation in Spanish. *Procesamiento del Lenguaje Natural*, (62):21–28.

Jiménez-Zafra, S. M., N. P. Cruz Díaz, R. Morante, y M. T. Martín-Valdivia. 2019b. NEGES 2019 Task: Negation in Spanish. En *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, CEUR Workshop Proceedings, Bilbao, Spain. CEUR-WS.

Jiménez-Zafra, S. M., N. P. Cruz-Díaz, M. Taboada, y M. T. Martín-Valdivia. 2020b. Negation detection for sentiment analysis: A case study in spanish. *Natural Language Engineering*, 1(1):1–30.

Jiménez-Zafra, S. M., M. T. Martín-Valdivia, L. A. U. Lopez, M. A. Martí, y M. Taulé. 2016. Problematic cases in the annotation of negation in spanish. En *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)*, páginas 42–48.

Jiménez-Zafra, S. M., R. Morante, E. Blanco, M. T. M. Valdivia, y L. A. U. Lopez. 2020a. Detecting negation cues and scopes in spanish. En *Proceedings of The 12th Language Resources and Evaluation Conference*, páginas 6902–6911.

Jiménez-Zafra, S. M., R. Morante, M. T. Martín-Valdivia, y L. A. U. Lopez. 2018a. A review of spanish corpora annotated with negation. En *Proceedings of the 27th International Conference on Computational Linguistics*, páginas 915–924.

Jiménez-Zafra, S., M. Taulé, M. Martín-Valdivia, L. A. Ureña-López, y M. A. Martí. 2018b. SFU ReviewSP-NEG: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns. *Language Resources and Evaluation*, 52(2):533–569.

Jimenez-Zafra, S. M., M. T. M. Valdivia, E. M. Camara, y L. A. Urena-Lopez. 2017. Studying the scope of negation for spanish sentiment analysis on twitter. *IEEE Transactions on Affective Computing*, 10(1):129–141.

Jiménez-Zafra, S. M., R. Morante, M. T. Martín-Valdivia, y L. A. Ureña-López. 2019c. Corpora Annotated with Negation: An Overview. *Computational Linguistics (Under review - Second round)*.

Martí, M. A., M. Taulé, M. Nofre, L. Marsó, M. T. Martín-Valdivia, y S. M. Jiménez-Zafra. 2016. La negación en español: análisis y tipología de patrones de negación. *Procesamiento del Lenguaje Natural*, (57):41–48.

Morante, R. y C. Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2):223–260.